The Ranking Algorithm of the Coach Browser for the UMLS Metathesaurus

Anna M. Harbourt, M.L.S.; Edmund J. Syed, B.S.; William T. Hole, M.D.;
Lawrence C. Kingsland, III, Ph.D.
Lister Hill National Center for Biomedical Communications
National Library of Medicine
Bethesda, MD 20894

ABSTRACT

This paper presents the novel ranking algorithm of the Coach Metathesaurus browser which is a major module of the Coach expert search refinement program. An example shows how the ranking algorithm can assist in creating a list of candidate terms useful in augmenting a suboptimal Grateful Med search of MEDLINE.

INTRODUCTION

Healthcare professionals are increasingly realizing the important role access to biomedical information can play in the patient care process [1,2]. Online databases constitute a significant source of this information. Computer systems that assist actively in the search refinement process can complement the functions in other programs in providing access to online information for those who need it.

The Coach expert search refinement program, a product of the Unified Medical Language System^{TM} (UMLS$^{®}$) initiative [3] at the National Library of Medicine (NLM), is such a system. Its goal is to apply the UMLS Knowledge Sources to help Grateful Med$^{®}$ users improve retrieval in MEDLINE$^{®}$ [4]. Coach works interactively with the user, with Grateful Med, with its knowledge sources and with NLM's ELHILL$^{®}$ mainframe retrieval engine [4,5]. The primary knowledge source Coach offers in helping users augment or otherwise improve their searches is the UMLS Metathesaurus$^{®}$, one of three knowledge sources currently available from NLM as a part of the Unified Medical Language System project [3].

Coach is one of several applications that are testing the use of the UMLS Metathesaurus to improve MEDLINE retrieval. An application developed at Yale University to augment Grateful Med searches locates a term in the Metathesaurus and identifies all synonyms and related terms of the term and continues the process recursively until twenty MeSH$^{®}$ terms are found [6]. The Physician's Information Assistant allows users to navigate through the schematic, semantic, and lexical relationships presented in Metacard, a Macintosh-based Metathesaurus browser, and to select terms to be used in a Grateful Med search [7]. The University of Pittsburgh's CHARTLINE system applies a lexical matching technique to find Metathesaurus terms present in full text patient records; those that are co-occurring MeSH terms are then presented to users as potential MEDLINE searches [8]. The approach employed in the Coach Metathesaurus browser differs from these applications in that it uses an algorithm to weight the selection of lexically similar terms occurring in Metathesaurus concept records.

COACH METATHESAURUS BROWSER

The 1993 Metathesaurus, Meta 1.3, includes concepts and terms from seventeen different vocabularies, including all of MeSH; all diseases in ICD-9-CM (the International Classification of Diseases, 9th edition, Clinical Modification); and a number of more specialized clinical vocabularies [9]. In the aggregate, the Metathesaurus contains hundreds of thousands of terms. Definitions; lexical variants; synonyms; related terms; co-occurrences of terms with other terms in articles indexed in MEDLINE; semantic type assignments; previous indexing for terms derived from MeSH; broader, narrower, hierarchical, and other relationships; and many other elements are present. Through its Metathesaurus browser module, the Coach program can map a user's term to related terms in MeSH and in other vocabularies. The Coach Metathesaurus browser is available as a standalone module in the 1993 release of the UMLS knowledge sources.

The Coach Metathesaurus browser's retrieval engine operates from a universe of 152,444 concepts, 311,046 terms (including lexical variants, synonyms and others). The total number of Metathesaurus data elements against which the browser operates is 799,173.

Since many of these entries are multi-word terms like "Unified Medical Language System", the total number of words involved is nearly one million: 937,920. Of

0195-4210/92/$5.00 © 1994 AMIA, Inc.

these hundreds of thousands of words, only those found to be unique after some processing become entries in the Coach browser's index file. The processing begins by discarding pluralizations and words of less than three characters. The remaining words are truncated at 10 characters, then duplicates are removed. The index file resulting from this process consists of 77,458 unique words or partial words of up to 10 characters.

RANKING ALGORITHM

To map from the user's input term to related terms in MeSH and other vocabularies, the Coach Metathesaurus browser uses a non-Boolean retrieval algorithm. It accepts a multiple-word term as input and produces a ranked list of Metathesaurus concepts as output. The weighting algorithm includes the inverse proportion of Metathesaurus concepts in which a query term hits [10]. For each word in a Metathesaurus query, the browser's retrieval engine searches multiple data elements of every concept record in the Metathesaurus. The list below shows the elements involved in the ranking algorithm and the powers-of-two weighting scheme they are assigned:

Meta heading:	32
Lexical variant:	16
Synonym:	8
Previously indexed:	4
Reviewed related:	2
Unreviewed related:	1

A score is calculated for every Metathesaurus concept in which the query term hits. The hits for each word of the query term in each Metathesaurus concept record are summed, then divided by the total number of occurrences of the query term across the entire Metathesaurus. This final inverse term proportion calculation helps give higher values to query terms that occur infrequently in the entire Metathesaurus but frequently in the individual Metathesaurus concept record which hit and is being scored. If a Metathesaurus query term contains more than one word (e.g., *erythema chronicum migrans*, a term with three words), the scores for the individual words which hit are summed. The sum is then multiplied by $25^{(N-1)}$, where N is the number of words in the term which hit. An example: one word hits, of a three-word term. The score for that word is multiplied by $25^{(0)}$, which is 1. Another example: two words hit, of a three-word term. The sum of the scores for those words is multipled by $25^{(1)}$, or 25. If all three words hit, the sum of their scores is multiplied by $25^{(2)}$, or 625. That three-word term will be ranked at the top or very high in the list of Metathesaurus concepts the browser returns.

Figure 1 is a pseudocode representation of the ranking algorithm of the Coach Metathesaurus browser's retrieval engine.

Figure 1: Pseudocode for Ranking Algorithm

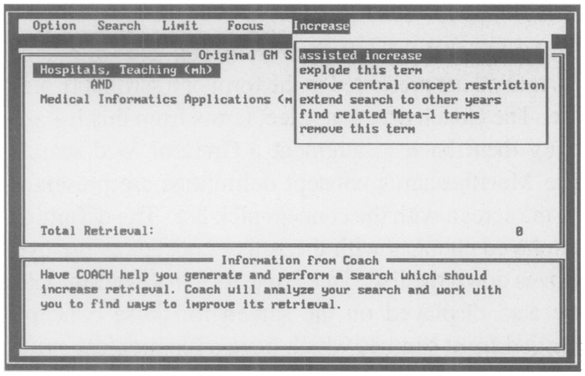
```
for all single words
{
 if (! lookup (word)) /* lookup was successful*/
 {
    read # of uids
    while (#--)
    {
       read uid, rank;
       makehash (uid);
       if (array [hashkey]>0)
        {
          array[hashkey] += (rank / # of terms);
          array[hashkey] *= 25;
        }
        else
        {
           array[hashkey] = (rank / # of terms);
         }
    }
 }
}
quicksort ();
display ();
```

Scores for each Metathesaurus concept are sorted and the ranked list presented in the form of a scrollable pick list. The Coach user can select terms from this list and bring them back to augment a Grateful Med search. The Metathesaurus concept definitions are presented on the screen with the concept pick list. The definition displayed changes with the active concept as the user moves down the list. Tree contexts, single or multiple, are also displayed on the screen for those concepts derived from sources which have a hierarchical structure, e.g., MeSH, CPT, and ICD-9-CM. The user can invoke a "Why this hit?" function to show in matrix form which weighted element for each word of a multiword query such as "HOSPITALS, TEACHING" caused a particular Metathesaurus concept to be retrieved.

This specific algorithm, which emphasizes MeSH, was chosen due to the role of the Coach Metathesaurus browser as a major component of the Coach expert search refinement system, which is designed to augment Grateful Med searches of MEDLINE. The data elements included in the ranking algorithm's weighting scheme were judged to best represent the MeSH perspective. As the Coach system moves towards the broader UMLS goal of retrieving information from heterogeneous databases beyond MEDLINE, the Coach Metathesaurus browser's ranking algorithm will have the ability to reflect the perspectives of multiple vocabularies towards which the Coach system may direct information queries. At present, if a concept originating from a vocabulary other than MeSH is selected, the Coach system searches the concept as textwords in MEDLINE through the Grateful Med search engine.

SEARCH EXAMPLE

The following example demonstrates the calculations involved as the ranking algorithm of the Coach Metathesaurus browser is applied against a group of hits in the Metathesaurus. The system is helping the user find additional Metathesaurus concepts which might improve a MEDLINE search. A Grateful Med search for articles discussing teaching hospitals that have implemented medical informatics applications produces no retrieval. The Coach search refinement system is invoked with one keypress. Coach parses the Grateful Med search. Since the search got less than 35 hits (an internal threshold in Coach), the program assumes the user wants more. Coach places the user's command highlight at its ASSISTED INCREASE menu command (Fig. 2).



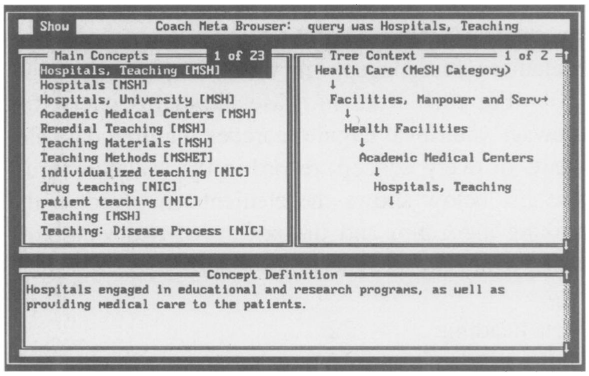
```
Option   Search   Limit   Focus   Increase
================ Original GM S assisted increase
Hospitals, Teaching (mh)       explode this term
      AND                      remove central concept restriction
Medical Informatics Applications (m extend search to other years
                               find related Meta-1 terms
                               remove this term

Total Retrieval:                                        0

=============== Information from Coach ===============
Have COACH help you generate and perform a search which should
increase retrieval. Coach will analyze your search and work with
you to find ways to improve its retrieval.
```

Figure 2: Coach ASSISTED INCREASE

The user presses Enter, invoking the ASSISTED INCREASE command Coach was offering. Coach begins a ten-step analysis of the terms of the original Grateful Med search, diagnosing specific problems for which it has fixes which should increase retrieval in MEDLINE. Coach's actions at this stage are based entirely on the nature and context of the terms it finds in the original Grateful Med search.

In this instance, Coach finds both the user's terms "explodable". It explodes both terms, which automatically incorporates into the search the several child terms for MEDICAL INFORMATICS APPLICATIONS (in the L1.700.508+ branch of the Information Science tree of the MeSH controlled vocabulary hierarchy, which includes such terms as INFORMATION SYSTEMS) and the single child term under HOSPITALS, TEACHING (which is HOSPITALS, UNIVERSITY).

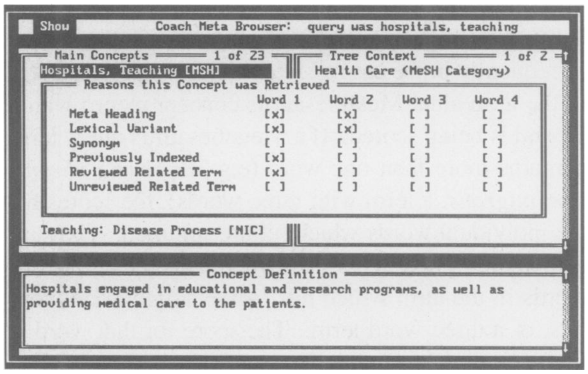
Next, Coach presents the user with displays from the Metathesaurus for each of the terms in the original Grateful Med search. The Coach browser creates the following display of related concepts from its Metathesaurus knowledge source when the argument HOSPITALS, TEACHING is passed to it (Fig. 3).



```
Show        Coach Meta Browser:  query was Hospitals, Teaching

== Main Concepts ====== 1 of 23 ==  == Tree Context ======== 1 of 2 =
Hospitals, Teaching [MSH]            Health Care (MeSH Category)
Hospitals [MSH]                        ↓
Hospitals, University [MSH]            Facilities, Manpower and Serv→
Academic Medical Centers [MSH]           ↓
Remedial Teaching [MSH]                  Health Facilities
Teaching Materials [MSH]                   ↓
Teaching Methods [MSHET]                   Academic Medical Centers
individualized teaching [NIC]                ↓
drug teaching [NIC]                          Hospitals, Teaching
patient teaching [NIC]                         ↓
Teaching [MSH]
Teaching: Disease Process [NIC]

==================== Concept Definition ====================
Hospitals engaged in educational and research programs, as well as
providing medical care to the patients.
```

Figure 3: Coach Meta Browser Concept Display

The first concept retrieved, HOSPITALS, TEACHING appears in the hit list because it hit in multiple elements of the Metathesaurus concept record. The first word (HOSPITALS) hit in the Meta Heading, Lexical Variant, Previously Indexed, and Reviewed Related data elements. The second word (TEACHING) hit in the Meta Heading and Lexical Variant data elements (Fig. 4).



```
Show        Coach Meta Browser:  query was hospitals, teaching

== Main Concepts ====== 1 of 23 ==  == Tree Context ======== 1 of 2 =
Hospitals, Teaching [MSH]            Health Care (MeSH Category)
  == Reasons this Concept was Retrieved ======================
                              Word 1    Word 2    Word 3    Word 4
     Meta Heading              [x]       [x]       [ ]       [ ]
     Lexical Variant           [x]       [x]       [ ]       [ ]
     Synonym                   [ ]       [ ]       [ ]       [ ]
     Previously Indexed        [x]       [ ]       [ ]       [ ]
     Reviewed Related Term     [x]       [ ]       [ ]       [ ]
     Unreviewed Related Term   [ ]       [ ]       [ ]       [ ]

  Teaching: Disease Process [NIC]

==================== Concept Definition ====================
Hospitals engaged in educational and research programs, as well as
providing medical care to the patients.
```

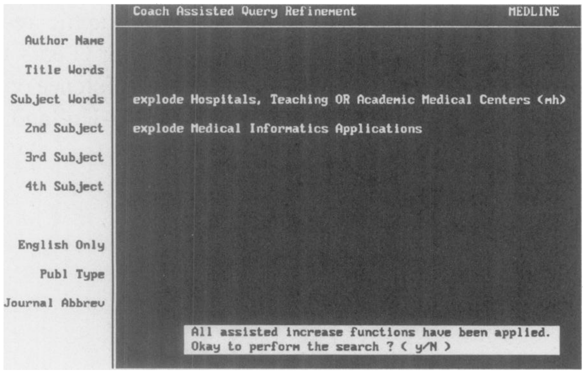
Figure 4: Coach Meta Browser Weighted Matrix Display

The weighting scores for the first word of the query term (HOSPITALS), by Metathesaurus concept data element in which it hit, are Meta heading (32); Lexical Variant (16); Previously Indexed (4); and Reviewed

Related Term (2). The scores are summed (54), then divided by the total number of the occurrences of the term HOSPITALS àcross the entire Metathesaurus (278). The final score for this word of the two-word term HOSPITALS, TEACHING is (54/278). The second word of the query term (TEACHING), hit in the Meta Heading and the Lexical Variant data elements; the second word's score (32+16) is divided by the total number of occurrences of the word TEACHING across the entire Metathesaurus (84). The final score for this word is (48/84). The scores for each word of the query term (0.194245 and 0.571429) are summed, then multiplied by 25 because the query HOSPITALS, TEACHING hit on two words. The final score for the concept retrieved is 19.141829.

Calculating the scores for the next several concepts gives the following results: HOSPITALS (5.091641); HOSPITALS, UNIVERSITY (5.091641); ACADEMIC MEDICAL CENTERS (1.079136). The scores for the next three concepts (REMEDIAL TEACHING, TEACHING MATERIALS, TEACHING METHODS) were the same, 0.571429, since the data elements in which they hit were identical. The scores for the remaining concepts in the display were also identical, 0.404762, since they all hit in the same data elements. When the calculated score of a group of concept hits is the same, those concepts are sorted using a quicksort algorithm.

The user selects the term ACADEMIC MEDICAL CENTERS from the pick list. Coach adds it to the users search. Coach creates a similar list of Metathesaurus concepts for the other Grateful Med search term: MEDICAL INFORMATICS APPLICATIONS. Coach then offers to submit the newly modified search strategy (Fig. 5) to NLM's ELHILL retrieval system through the Grateful Med search engine®.



Coach Assisted Query Refinement MEDLINE

Author Name

Title Words

Subject Words explode Hospitals, Teaching OR Academic Medical Centers (mh)

2nd Subject explode Medical Informatics Applications

3rd Subject

4th Subject

English Only

Publ Type

Journal Abbrev

All assisted increase functions have been applied.
Okay to perform the search ? (y/N)

Figure 5: Coach Assisted Query Refinement Screen

The modified search retrieves 105 hits. At this point, a user might choose to download some of the records or might reinvoke Coach. Coach, noting that the retrieval was above its 35-hit threshold, assumes the user probably wants fewer hits. It places the user's command highlight at its ASSISTED FOCUS menu command. The user does indeed want less retrieval, so Coach begins an analysis of the terms of the modified search strategy looking for ways to focus retrieval more tightly. For each term of the search, Coach asks the user whether the "central concept" restriction should be applied. This restricts retrieval to citations in which a human indexer has considered this term to be a major topic or central concept of the article. Applying this central concept restriction to all the terms in the search and resubmitting the search to ELHILL results in fifteen good hits. Several are shown below:

TI - Prototyping an institutional IAIMS/UMLS information environment for an academic medical center.

*MH - *Academic Medical Centers*

MH - Computer Communication Networks

MH - Computer Systems

MH - Databases, Bibliographic

MH - Databases, Factual

MH - Information Storage and Retrieval

*MH - *Integrated Academic Information Management Systems*

MH - National Library of Medicine (U.S.)

MH - Support, U.S. Gov't, P.H.S.

*MH - *Unified Medical Language System*

MH - United States

TI - Successful principles for collaboration formation of the IAIMS consortium.

*MH - Academic Medical Centers/*ORGANIZATION & ADMIN*

MH - Computer Communication Networks/*ORGANIZATION & ADMIN

MH - Human

*MH - *Information Systems*

MH - *Interinstitutional Relations

MH - National Library of Medicine (U.S.)

MH - Support, U.S. Gov't, P.H.S.

MH - United States

DISCUSSION

The Coach Metathesaurus browser created, ranked and returned a hit list of related concepts far richer than that which would be returned by simple alphabetic term permutation or Boolean search. The Coach browser expands the semantic context of a search term to include additional terms connected through the complex infrastructure of terms in the ranking algorithm: lexical variants, synonyms, previously indexed, reviewed related and unreviewed related terms. In the simple example above, the additional search term ACADEMIC MEDICAL CENTERS, added to the search as a component of Coach's interactive search refinement process, resulted in improved MEDLINE retrieval.

In the future, when the goal is to increase retrieval and a user has selected the parent term of a child already contained in the strategy, Coach will know to explode the parent term and delete the child term. This will be a further refinement of the search example above, since ACADEMIC MEDICAL CENTERS is the parent term of HOSPITALS, TEACHING -- the user's original term in the initial Grateful Med search. Coach will also caution users to be judicious in the application of the central concept restriction to search terms.

The effectiveness of the current algorithm of the Coach Metathesaurus browser has yet to be tested experimentally. The Coach Expert Search Refinement System is in the product development stage and is being beta tested with a variety of external NLM collaborators . As a part of that process, feedback concerning the effectiveness of the Coach Metathesaurus browser and thė usefulness of the algorithm will be collected. The development and testing phase will also include controlled experiments involving the UMLS test collection retrievals [11] as a benchmark to which retrieval resulting from the Coach Metathesaurus browser intervention can be compared.

Reference

[1] Siegel ER, Lindberg DAB, Rapp BA, Wilson SR. Evaluating the impact of MEDLINE using the Critical Incident Technique. In: Lun KC, Degoulet P, Piemme TE, Rienhoff O, ed(s). MEDINFO 92. Proceedings of the Seventh World Congress on Medical Informatics. Amsterdam: North-Holland, 1992:1287-92.

[2] Haynes RB, McKibbon KA, Walker CJ, Ryan N et al. Online access to MEDLINE in clinical settings. A study of use and usefulness. Ann Intern Med 1990 Jan 1;112(1):78-84.

[3] Lindberg DAB, Humphreys BL, McCray AT. The Unified Medical Language System. Methods Inf Med 1993 July;32(4):in press.

[4] Kingsland LC III, Syed EJ, Lindberg DAB. Coach: an expert searcher program to assist Grateful Med users searching MEDLINE. In: Luñ KC, Degoulet P, Piemme TE, Rienhoff O, ed(s). MEDINFO 92. Proceedings of the Seventh World Congress on Medical Informatics. Amsterdam: North-Holland, 1992:382-6.

[5] Kingsland LC III, Harbourt AM, Syed EJ, Schuyler PL. Coach: applying UMLS knowledge sources in an expert searcher environment. Bull Med Libr Assoc 1993 Apr;81(2):178-83.

[6] Jachna JS, Powsner SM, Miller PL. Augmenting Grateful Med with the UMLS Metathesaurus: an initial evaluation. Bull Med Libr Assoc 1993 Jan;81(1):20-8.

[7] Nelson SJ, et al. The Physician's Information Assistant. In: Clayton PD, ed. Proceedings of the 15th Annual Symposium on Computer Applications in Medical Care. New York: McGraw-Hill, 1992:950-52.

[8] Miller RA, Gieszczykiewicz FM, Vries JI, Cooper GF. CHARTLINE: providing bibliographic references relevant to patient charts using UMLS Metathesaurus Knowledge Sources. In: Frisse ME, ed. Proceedings of the 16th Annual Symposium on Computer Applications in Medical Care. New York: McGraw Hill, 1993:86-90.

[9] National Library of Medicine Fact Sheet. UMLS Metathesaurus. National Institutes of Health, July, 1993.

[10] Sparck Jones K. A statistical interpretation of term specificity and its application in retrieval. J Documentation 1972;28:11-21.

[11] Schuyler PL, McCray AT, Schoolman HM. A test collection for experimentation in bibliographic retrieval. In: Barber B, Cao D, Qin D. Wagner G ed(s.) MEDINFO'89: Proceedings of the Sixth Conference on Medical Informatics. Amsterdam: North-Holland, 1989;910-2.